International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning for Tax Optimization: Analyzing Chargeable Income and Property Tax

Ritika S. Kedar

PG Student, Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Tax optimization is important in financial planning and policy-making helping to secure revenue collection and economic growth. To categorize taxable entities and predict tax collections, the analysis uses machine learning models like Decision Tree Classifiers and Regressors together with statistical techniques like correlation analysis and feature engineering. Information designs, deviations, and patterns can be recognized utilizing visualization instruments such as warm and center thickness gauges (KDE). The comes about appear noteworthy contrasts within the conveyance of Charactive salary, centering on the effect of charge arrangement and financial alter. Changes in property tax also show that different types of properties contribute different amounts of income, hence the need for data-driven tax initiatives. This result shows that Machine learning can improve tax calculation, tax predictions, provide insights for better revenue forecasting, tax policy optimization decision-making and provide personalized tax recommendations. This research also focuses on how artificial intelligence can make tax planning more efficient and exact in Singapore's evolving tax system.

KEYWORDS: *ML, Tax Optimization, Predictive Analytics, Data-Driven Tax Strategies.*

I. INTRODUCTION

Tax optimization plays an imperative part in monetary arranging of both businesses and people, encouraging compliance whereas overseeing money related commitments inside the bounds of lawful controls [10]. In Singapore, recognized for its favorable assess environment, viable charge arranging is basic to completely advantage from the motivating forces and findings accessible [1]. However, conventional tax optimization techniques typically depend on rule-based systems and manual calculations, which can be labor-intensive and susceptible to inefficiencies [5]. The advent of machine learning (ML) presents a revolutionary method for tax optimization, utilizing data-driven models to scrutinize financial records, uncover tax-saving prospects, and enhance decision-making processes [6]. By employing ML algorithms, tax professionals and organizations can streamline the evaluation of chargeable income and property tax, thereby achieving improved financial results with enhanced precision and efficiency [13].

This research article investigates the utilization of machine learning (ML) in optimizing taxation, with a particular focus on the assessment of chargeable income and property tax calculations in Singapore. This attention is sought by analyzing how charging information can be processed that can be checked, assessable amounts, and technology to evaluate investment funds, which can recommend technology to evaluate investment funds, highlighting the possibility of AI-controlled arrangements in modern assessments [15]. Furthermore, this paper deals with notable challenges such as information security, administrative obstacles, and the interpretability of the ML model in reviews [12].

Finally, the Extended Connection and Taxation Survey includes ways in which ML adheres to fee fees, minimizes errors and provides important experiences to both Singaporean citizens and political decision makers.

II. RELATED WORK

Previous research has mostly focused on fraud detection, corporate tax optimization, and property tax valuation. In order to improve regulations, identify fraud, and minimize tax obligations [1],[2],[13]. Machine learning has been used more and more in the tax industry. Almost all researchers use machine learning and artificial intelligence for tax optimization [20],[6]. Predictive modeling for tax compliance or tax liabilities has been studied in the past using supervised learning methods such as decision trees and neural networks to evaluate taxpayer behavior. Or we can say that researchers used these techniques to study how people pay taxes and fraud detection automatically [3],[12]. Similarly, [12] demonstrated how to use machine learning to automate fraud recognition and control. This reduces manual workers and at the same time improves accuracy. Various studies focus on tax optimization, particularly with regard to integrated income [11]. To allow for the best application of business tax planning approaches, suggested a deep learning model that can find incentives plus deductions required to reduce taxable income [5]. The e-signature tracking of management authority in artificial intelligence and blockchain technology has been the most popular choice for the companies operating in the business.

Overall, most research focused on the data that is already available. The majority of tax optimization research is mostly based on historical data. So, I focused on the real-time data to create a model that shows the changes in Singapore's economy and tax laws.

III. DATA AND SOURCES OF DATA

Data:

The FIN Tech project in Singapore is based on an analysis of two important financial data sets:

1. **Chargeable Income (2003–2018)** - This data record provides insight into taxable income trends over a 15-year period, determines distributions and determines the impact of tax payer tax distributions. You can determine the identification of a pattern.
2. **Property Tax Collection** - This data record contains property tax revenue records that include valuable

information about property taxes, property trends and government income from tax taxes.

Data Sources:

1. **RAS (Revenue and Administrative Sources) Income Statistics:**

IRAS (Inland Revenue Authority of Singapore) publishes comprehensive profit and loss statistics, including data on easy income over the years.

2. **Singapore Open Data Portal:**

The government's open data portal hosts data records in relation to income statistics. For example, the data record "Company Chargeable Income by Tax Group" provides detailed information on corporate chargeable income.

3. **IRAS Basic Tax Statistics:**

The IRA provides detailed statistics on real estate taxes that you can access here. Real estate types from fiscal year 2002 (HDB, private housing, trade, industry, etc.).

IV. RESEARCH METHODOLOGY

Research Methodology for Machine Learning in Tax Optimization: Chargeable Income and Property Tax in Singapore are analyzed using a systematic, data-driven approach. The process begins with data collection. It requires acquiring information from a variety of sources. The sources include reports from Singapore's Inland Revenue Authority (IRAS), real estate market statistics, business tax filings, and economic indices like GDP and inflation rates. A comprehensive dataset is ensured by taking into account both primary and secondary data sources. The dataset shown in the figure below includes information such as the year of assessment, tax group, total income, chargeable income, group relief, etc.

```
[2]: df = pd.read_csv('../../DataSets/Fin_singa/Income-comp_annual/chargeable_income_tax_y.csv')

[3]: df
```

t[3]:

oup_relief	loss_carryback_relief	ci_aft_reliefs	tax_exemption	ci_aft_exemption	gross_tax_payable	tax_deducted_at_source	other_tax_set_offs	net_tax_assessed
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
na	na	na	na	na	na	na	na	na
375948	6841	38762221	363999	35122222	388157	413751	28489	-54083
630173	20259	64328862	1341043	62987819	10670606	1002273	223336	9444997

Fig.1 Dataset: Chargeable Income and Tax Collection

```
In [2]: df = pd.read_csv('../../DataSets/Fin_singa/Income-comp_annual/chargeable_income_tax_y.csv')

In [3]: df

Out[3]:
```

	year_of_assessment	tax_group	no_of_companies_assessed	total_income	donations	chargeable_income	group_relief	loss_carryback_relief	ci_aft_reliefs
0	2003	Non-Taxable Group	78265	-15795068	3035	-15798091	na	na	na
1	2003	Taxable Group	28530	-44426364	17165	-44481160	na	na	na
2	2004	Non-Taxable Group	77484	-22026713	61069	-22087766	na	na	na
3	2004	Taxable Group	29366	36206700	19524	36187176	na	na	na
4	2005	Non-Taxable Group	80653	-20406048	76668	-20482716	na	na	na
5	2005	Taxable Group	29595	45318884	18468	45287804	na	na	na
6	2006	Non-Taxable Group	82535	-22186440	63207	-22249957	na	na	na
7	2006	Taxable Group	32032	49259719	36626	49223093	na	na	na
8	2007	Non-Taxable Group	87167	-33507663	64068	-33571731	3369.48	6841	38762221
9	2007	Taxable Group	36401	65060196	80672	64979524	630173	20259	64328862

Fig.2 Dataset: Chargeable Income and Tax Collection

Figure 1 and Figure 2:

1. Data Overview:

The dataset comprises of data approximately chargeable salary assess for diverse a long time of evaluation. It incorporates highlights such as year of appraisal, charge gather, number of companies evaluated, add up to wage, gifts, chargeable pay, gather help, misfortune carry back alleviation, chargeable pay after reliefs, assess exception, chargeable pay after exception, net charge payable, charge deducted at source, other assess set offs and net assess surveyed.

2. Data Collection and Preprocessing:

Information about payroll assessments is stacked from CSV datasets using Pandas. Libraries such as NumPy and Communications were imported separately for numerical manipulation and Communications handling.

3. Data cleaning procedures:

This handles the lost values by replacing "Na" with NAN and changing it via the Float64 column. Evacuating negative signs from the 'chargeable income' column utilizing outright esteem work.

4. Exploratory Data Examination (EDA):

- Information outline data and dtypes were investigated utilizing df.info () and df. dtypes.
- Esteem checks were gotten for the 'chargeable income' column to get it pay dispersion. work shown the primary few rows of the information to urge a set of the information.
- Column names were recovered utilizing df. Columns.

5. Feature Engineering:

Modern highlights were made based on existing ones:

- ci_aft_reliefs: Chargeable wage after applying bunch alleviation and misfortune carryback alleviation.
- gross_tax_payable: Net assess payable calculated as a rate of ci_aft_exemption.

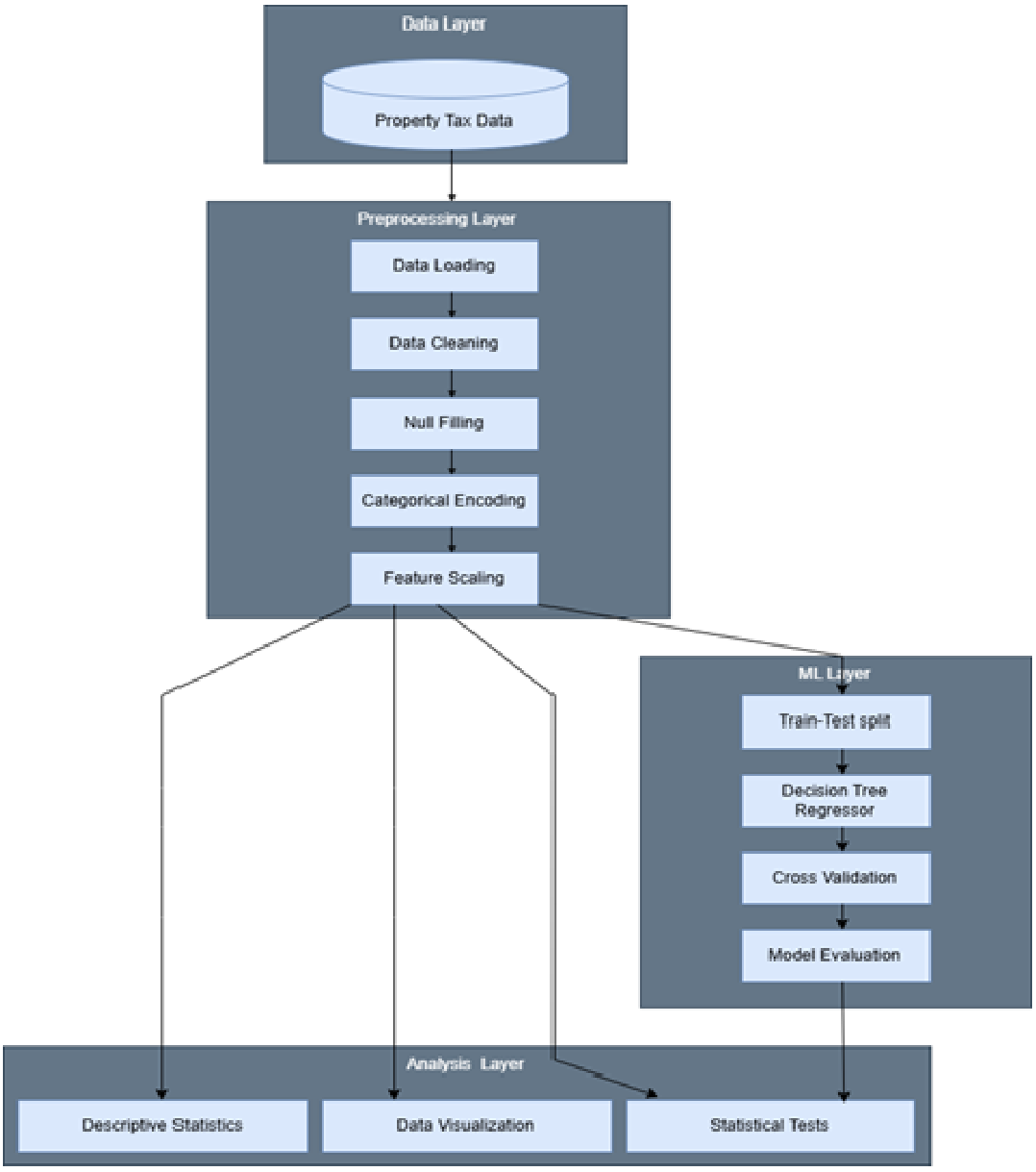

Fig.3: Flowchart

Figure 3:

1. **Data Loading:** Data Loaded using Python's Pandas library in Jupyter Notebook.
2. **Data Preprocessing:** Filled null values using Median values according to the property type. Before applying ML model the categorical data was converted into numerical data using pd.getdummies() function of pandas library.
3. **Train-Test split:** Train tests have been conducted. 80% of the data was used to train the model and had 20% tested performance. This is a standard practice for assessing how well a model is generalized to invisible data.
4. **Model Training:** We decided on a tree regressor as a model for machine learning and chose to train with training data. Decision trees can contain complex relationships in your data and are suitable for regression tasks such predicting property tax amounts.
5. **Model Evaluation:** R-squared metrics (coefficient of determination) and medium-directional errors (MSE) were used to evaluate model Performance.
6. **R-Squared:** measures the variance sharing of target variables declared by the model.
7. **MSE:** Model, measures the percentage of variance interpreting these metrics in the context of the problem domain.
8. **Feature Scaling:** Standardizes both input functions and target variables using standard scales.
9. **Cross-validation:** To assess the performance of the model, cross-validation was performed at CV = 10. Cross-validation provides a more reliable estimate of the power of the model by evaluating it on multiple subsets of data.
10. **Model Optimization:** I installed a decision tree regressor model for the entire data record after scaling. This step is essential for providing a model to predict invisible data.
11. **Final Evaluation:** The performance of the optimized models was evaluated using the same metrics (MSE and R Square) via the dataset. This step provides a comprehensive assessment of the effectiveness of the model.
12. **Data Analysis:** To get meaningful statistical insights, visualization and apply ML model using python libraries such as Pandas, Seaborn, Matplotlib, and Sklearn.

V. RESULTS AND DISCUSSION

Data Analysis and Exploration

- **Descriptive Statistics:**

This data shows an increasing trend in Property transactions and tax collection in Singapore from 2002 to 2019, with over 1 million cases. While the mean annual property value is 6.68 million, but a much lower median (79,843) suggests that some high-quality properties slow the average. Similarly, property tax collection varies widely depending on the average of 579,409, but extreme values reaching 1.71 million indicate unequal tax fees.

	financial_year	no_of_cases	annual_value	median_annual_value	property_tax_collection
count	90.000000	9.000000e+01	3.000000e+01	50.000000	9.000000e+01
mean	2010.500000	2.539732e+05	6.859096e+06	79483.400000	5.794093e+05
std	5.217193	3.271393e+05	2.107147e+06	98361.330829	4.487204e+05
min	2002.000000	1.019000e+03	3.991085e+06	8100.000000	3.238200e+04
25%	2006.000000	2.840750e+04	5.402162e+06	24000.000000	2.337702e+05
50%	2010.500000	7.712300e+04	6.472180e+06	39750.000000	4.557800e+05
75%	2015.000000	3.334135e+05	8.228930e+06	60500.000000	8.385215e+05
max	2019.000000	1.002671e+06	1.263014e+07	335000.000000	1.715165e+06

Fig 4: Mean, Median, Standard Deviation

In Figure 5, the results show the percentage of cases and property tax collection, but the annual and average annual values include missing values (NAN). The number of cases shows significant variation following a wide distribution of 28,407 (25th percentile) and 333.413 (75th percentile). Similarly, property tax collections show a wide range of 455,780. The missing rates for annual and median annual values indicate incomplete or inconsistent data that may affect accurate analysis.

```
Percentile values of no_of_cases
 25th Percentile: 28407.5 , 50th Percentile: 77123.0 , 75th Percentile: 333413.5

Percentile values of Annual Value
 25th Percentile: nan , 50th Percentile: nan , 75th Percentile: nan

Percentile values of Median Annual Value
 25th Percentile: nan , 50th Percentile: nan , 75th Percentile: nan

Percentile values of Property Tax Collection
 25th Percentile: 233770.25 , 50th Percentile: 455780.0 , 75th Percentile: 838521.5
```

Fig 5: Percentile

In Figure 6, the comes about appear relationship values between distinctive monetary measures. The middle yearly esteem (0.990279) has the most noteworthy positive relationship. This demonstrates a solid relationship with diverse components and the probability of yearly values. Yearly esteem (0.351697) too appears a decently positive relationship, demonstrating that this contributes to common money related patterns. In differentiate, there's a negative relationship between property charge collection (-0,262027) and case tally (-0,071242), which infers an reverse relationship.

no_of_cases	-0.071242
annual_value	0.351697
median_annual_value	0.990279
property_tax_collection	-0.262027

Fig 6: Kurtosis

- **Data Visualization:**

The density plot (Figure 7) shows the distribution of number of cases, annual values, median annual value, and property tax collection. This indicates that the majority of the values are concentrated on low values and some high-quality outliers.

- Statistical Test: After using the ML model, the Chi Square test was run.
- Key insights: Agreeing to the KDE chart, information is watched to be skewed to the correct. The data contains as it were a couple of exceptions. Distinguished in a box plot.

The visualization of the required boxplot gives understanding into the dissemination and spread of four critical factors. Cases, annual value, medium annual value, property tax collection. All four box plots, especially the presence of outliers in the middle annual values and annual values, show a skewed distribution with extreme values that affect the data distribution. Despite these variations, a chi-square value of 0.0 shows the value without correctly predicting that the ML model is optimally performed and that there is no deviation between the observed and expected results.

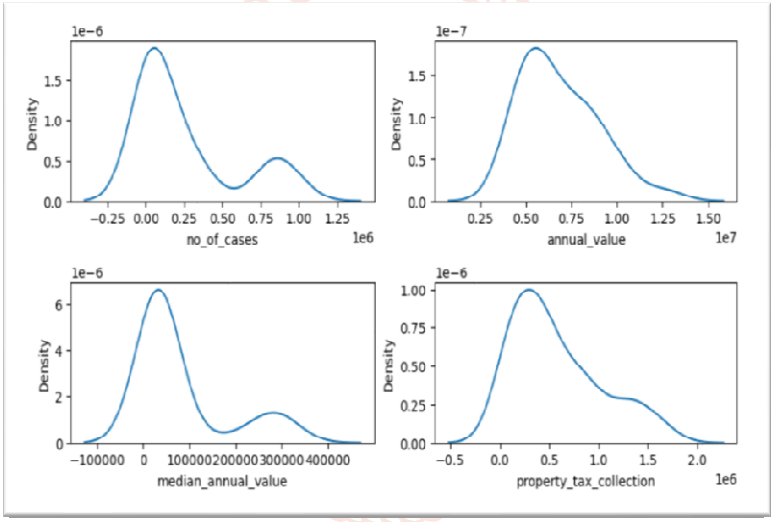

Fig 7. KDE Plot

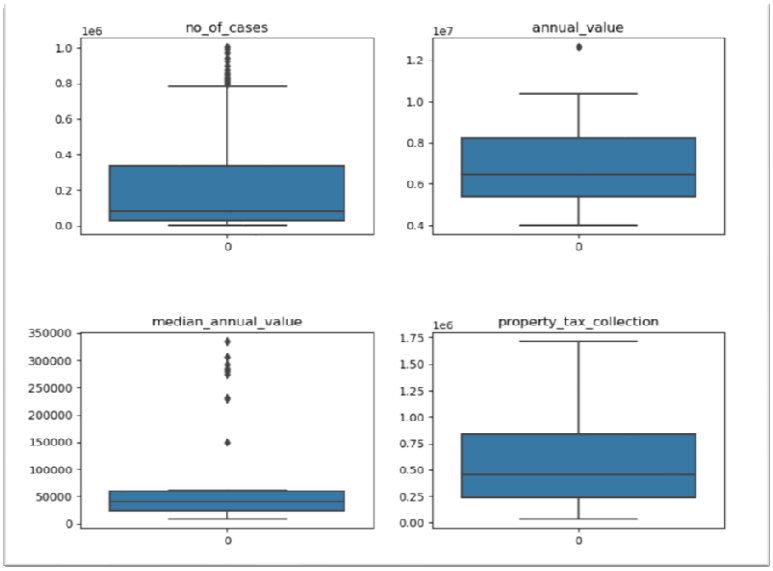

Figure 8: Boxplot Representations of No. of Cases, Annual Value, Median Annual Value, and Property Tax Collection

VI. CONCLUSION

This investigate underscores the critical part of machine learning in revolutionizing assess optimization, especially within the examination of chargeable salary and property charge in Singapore. By leveraging machine learning models such as decision tree regressors, beside factual procedures like relationship investigation and include building, this think about illustrates the adequacy of data-driven approaches in moving forward charge appraisal, income determining, and approach detailing.

The comes about demonstrate that machine learning improves the exactness of charge forecasts by recognizing information designs, inconsistencies, and patterns, subsequently minimizing mistakes and guaranteeing way better compliance.

Moreover, visualization methods such as KDE plots and box plots give more profound experiences into the dissemination of monetary factors, highlighting incongruities in assess commitments over diverse property sorts and pay bunches. The think about too uncovers that charge approach changes altogether affect wage conveyance and income collection, emphasizing the require for versatile, real-time charge methodologies. In spite of the promising comes about, challenges such as information security concerns, administrative impediments, and the interpretability of machine learning models stay. Tending to these challenges requires the integration of progressed AI strategies, real-time information preparing, and blockchain-based assess following to upgrade straightforwardness and effectiveness in charge organization.

Eventually, this investigate strengthens the potential of manufactured insights in present day assess frameworks, clearing the way for mechanized, data-driven, and impartial charge approaches that can advantage both taxpayers and governments. Future ponders ought to investigate the utilize of profound learning models and cross breed AI systems to assist optimize charge planning and financial determining.

VII. REFERENCES

[1] Agarwal, S., Qian, W., Yeung, B., & Zheng, H. (2023). Taxing the rich to finance redistribution: Evidence from a permanent tax increase in Singapore. *Management Science*.

[2] Alm, J., & Sennoga, E. (2010). Mobility, competition, and the distributional effects of tax evasion. *National Tax Journal*, 63(4), 1055-1084.

[3] Bian, T., Chen, J., Feng, Q., & Li, J. (2020). Comparing econometric analyses with machine learning approaches: A study on Singapore private property market. *Singapore Economic Review*, 1-24.

[4] Brueckner, J. K., & Kim, H. A. (2003). Urban sprawl and the property tax. *International Tax and Public Finance*, 10(1), 5-23.

[5] Chetty, R., Friedman, J. N., & Saez, E. (2013). Using differences in knowledge across neighbourhoods to uncover the impacts of the EITC on earnings. *American Economic Review*, 103(7), 2683-2721.

[6] Górski, Ł., Kuźniacki, B., Almada, M., Tyliński, K., Calvo, M., Asnaghi, P. M., ... & Nigrelli, J. I. (2024). Exploring explainable AI in the tax domain. *Artificial Intelligence and Law*.

[7] Hull, I., & Grodecka-Messi, A. (2022). Measuring the impact of taxes and public services on property values: A double machine learning approach. *arXiv preprint arXiv:2203.14751*.

[8] Kwa, H. L. D. (2019). Property tax in Singapore: Progressiveness of the regime on residential properties and its revenue potential. *ScholarBank@NUS Repository*.

[9] Oates, W. E. (1969). The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy*, 77(6), 957-971.

[10] Ooi, V. (2021). Tax law and the digital economy in Singapore. *SMU Centre for AI & Data Governance Research Paper*, 1-23.

[11] Piketty, T., Saez, E., & Stantcheva, S. (2014). Optimal taxation of top labor incomes: A tale of three elasticities. *American Economic Journal: Economic Policy*, 6(1), 230-271.

[12] Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2021). Tax evasion risk management using a hybrid unsupervised outlier detection method. *arXiv preprint arXiv:2103.01033*.

[13] Wong, C. H., & La, N. (2024). Applying machine learning in tax revenue forecasting. *Victoria's Economic Bulletin*,8(2).

[14] Zeng, T., & Zhang, Y. (2011). The impact of property taxes on income distribution across municipalities. *Public Finance Review*, 39(5), 645-676.

[15] Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*.